Research Paper | Computer Science | E-ISSN No : 2454-9916 | Volume : 3 | Issue : 5 | May 2017

AN ENSEMBLE APPROACH FOR CLASSIFICATION OF THYROID DISEASE WITH FEATURE OPTIMIZATION

A. K. Shrivas [1] | Pallavi Ambastha [1]

[1] Dept. Of IT, Dr. C. V. Raman University, Bilaspur (C. G.), India.

ABSTRACT

Now a day's people are facing lots of problem related to health. Diseases are also increasing due to increase number of populations. The survey helps to identify how the data mining techniques predict the thyroid disorder at earlier stage. Classification techniques play very important role to identify the disease in medical data. In this paper, the main objective is to classify the data as thyroid or non-thyroid and improve the classification accuracy. We have proposed robust ensemble model using various classification techniques like random forest, Naïve Bayes and K-Nearest Neighbors (K-NN). The proposed model gives better classification accuracy as 93.55%.We have also applied the feature optimization technique that is optimized selection to eliminate the irrelevant feature from data set and computationally improve the performance of model .The proposed model achieved better classification technique as 97.61% of accuracy with reduced 3 feature subset.

KEYWORDS: Thyroid, Decision tree, Classification, Ensemble Model, Optimization Selection.

I. INTRODUCTION

In medical science, diagnosis of health condition is very challenging task. Peoples are facing various health disease problems in which thyroid is very critical problem faced by the human being. Thyroid decease classification is one of the important problem in medical science because it is directly related to health of human body, these type of decease can be solve by proper and carefully treatments. A modern medical diagnosis system based on decision based system and find the problem based on classification of data. The main purpose of this work is to study about the Thyroid Disease with the help of Data Mining Techniques. Data mining plays a vital role in medical field for diagnosis of disease. It offers lot of classification techniques to predict the disease accuracy.

There are various authors have worked for classification of thyroid disease. S. Gaikwad et.al. [1] have suggested random forest for classification of thyroid data. The suggested model gives 96.63% of accuracy. A. Upadhyay, et al. [2] have used two decision tree classifier as C4.5 and C5.0 for classification of thyroid disease. C5.0 model gives 95% of accuracy which is better than C4.5 classifier N. Sigh [3] has suggested Support Vector Machine (SVM) is better classifier as compared to K-NN and Bayesian Net. Accuracy of SVM gives 84.62% of accuracy. M. C. Frates [4] suggested different image classifiers are Artificial Neural Networks (ANN), Support Vector Machines (SVM), Fuzzy measures, Genetic Algorithms (GA), Fuzzy support Vector Machines (FSVM) for classification of thyroid disease. The textural features in ANN help to resolve misclassification. SVM is the best available machine learning algorithms in classifying high-dimensional data sets. . D. Kerana Hanirex et al. [7] have suggested NNge model for classification of thyroid disease. NNge classifier gives 96.44% of accuracy with reduced number of features. Lavanya, D., et al. [8] have suggested CART classifier and compared with other decision tree classifier as C4.5 and ID3 for classification of thyroid data. The CART achieved highest accuracy as 94.68% as best model. S. Panday et al. [12] have used various classifiers like C4.5, Random Forest, Multilayer preceptor and Bayes Net for classification of thyroid data. The classifier C4.5 gives better classification accuracy compare to others. K.Geeta et al. (2016) [13] have propped Evolutionary Multivariate Bayesian prediction classifier for classification of thyroid disease. K. Rajam [14] has discussed the use of data mining techniques for classification of thyroid disease and specially explore as Naïve bayes, decision tree, back propagation, support vector machine in the context of thyroid disease.

In this paper various data mining techniques like Random Forest, Naïve Bayes and K-NN are used to develop classifier for diagnosis and classification of thyroid disease. A data set downloaded from UCI repository site is used for the experimental purpose, entire work is carried out with Rapid Miner Studio software under Windows 7 environment.

II. DATA SET DESCRIPTION

Thyroid dataset is taken from UCI machine learning repository [5]. Dataset is given from Garavan institute and documentation is given by Ross Quinlan. Database consists of patients records. Each record is having 29 features, 7547 instances and 1 class having thyroid and non thyroid. Features are boolean or continuous valued. The features are namely Age, Sex, On thyroxine, Query on thyroxine, On antithyroid medication, Sick, Pregnant, Thyroid surgery, Il1treatment , Query hypothyroid, Query hyperthyroid, Lithium, Goitre, Tumor, Hypo pituitary, Psych, TSH measured, ,TSH, T3 measured, T3, TT4 measured, TT4, T4U measured, T4U, FTI measured, FTI, TBG measured, TBG and Referral source.

III. CLASSIFICATION TECHNIQUES

- Decision tree [6] is very popular data mining technique. A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node. In this research work we have used Random forest for classification thyroid data.

  Random Forest (or RF) [9] is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the classes output by individual trees. Random Forests are often used when we have very large training datasets and a very large number of input variables (hundreds or even thousands of input variables). A random forest model is typically made up of tens or hundreds of decision trees.

- Bayesian classification [6] is based on Bayes' Theorem. Bayesian classifiers are the statistical classifiers. Bayesian classifiers can predict class membership probabilities such as the probability that a given tuple belongs to a particular class. Classification algorithms have found a simple Bayesian classifier known as naïve Bayes classifier to be comparable in performance with decision tree and selected neural network classifiers. Bayesian classifiers have also exhibited high accuracy and speed when applied to large dataset.

- The k-nearest-neighbor method [6] was first described in the early 1950s. The method is labor intensive when given large training sets; It has since been widely used in the area of pattern recognition. Nearest-neighbor classifiers are based on learning by analogy, that is, by comparing a given test tuple with training tuples that are similar to it.

IV. FEATURE OPTIMIZATION AND ENSEMBLE MODEL

- Feature selection [10] is an optimization process in which one tries to find the best feature subset from the fixed set of the original features, according to a given processing goal and feature selection criteria. A solution of an optimal feature selection does not need to be unique. Different subset of original features may guarantee accomplishing the same goal with the same performance measure. An optimal feature set will depend on data, processing goal, and the selection criteria being used.

  In this research work, we have used optimization selection technique is used to optimize the original feature set. This approach is used two deterministic greedy feature selection algorithms forward selection and backward elimination are used for feature selection [15].

- An ensemble model [11] combines the output of several classifier produced by weak learner into a single composite classification. It can be used to reduce the error of any weak learning algorithm. The purpose of combining all these classifier together is to build a hybrid model which will improve classification accuracy as compared to each individual classifier.

V. RESULTS AND DISCUSSION

This research work done in Rapid miner data mining tools in window environment. We have used various classification techniques like random Forest, Naïve Bayes and K-NN for classification of thyroid disease. We have applied the thyroid data set into classification techniques with 70-30% training-testing partition. Individual's models are not giving satisfactory results. We have proposed

new ensemble model that is combination of Random Forest, Naïve bayes and K-NN which gives better classification accuracy as 93.55% compare to other individuals models. Table 1 shows that accuracy of individuals and proposed ensemble model. Fig.1 shows that graphical representation of confusion matrix of proposed ensemble model. The confusion matrix can be used to calculate the performance of models. Table 2 shows that performance measures of proposed ensemble model like sensitivity, specificity and accuracy to check the robustness of models.

Feature optimization is optimizing the feature from original feature space. In this research work we have used optimization selection to optimize the feature subset and increase the computational time and accuracy of model. Table 3 shows accuracy of proposed ensemble model with feature optimization technique. Our proposed ensemble model achieved better accuracy as 97.61% with 3 numbers of features. Finally our proposed ensemble model is better for classifying thyroid and non thyroid disease with high accuracy and less computation time.

Table 1: Accuracy of models with 70-30% training –testing data partition

SN	Individual Model Name	Accuracy (%)
2.	Random Forest	91.39%
5.	Naïve Bayes	92.67%
6.	K-NN	90.06%
7	Random Forest+ Naïve Bayes+ K-NN	93.55%

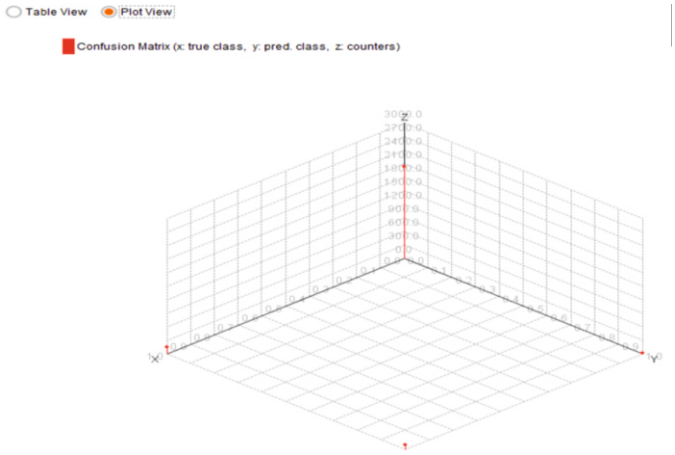

Fig 1: Graphical representation of confusion matrix of proposed model

Table 2: Performance measures of proposed ensemble model

Accuracy	93.49%
Sensitivity	93.55%
Specificity	91.78%

Table 3: Feature optimization technique on proposed ensemble model

Feature selection technique	Number of features	Name of features	Accuracy (%) after feature selection
Optimize Selection (forward & backward elimination)	3	TSH, T3 measured , T3	97.61

VI. CONCLUSION

Identification of disease is very critical problem in medical science. Classification is one of the important techniques to classify the data as thyroid or non-thyroid disease. Various research works have done in the field of thyroid classification and different data mining techniques used to build robust classifiers. In this paper, new proposed ensemble model is developed for classification of thyroid disease with high accuracy. We have also applied the feature optimization technique to computationally increase the performance of model. Our proposed model gives satisfactory result as 97.61% of accuracy with few numbers of features.

REFERENCES

1. S. Gaikwad and N. Pise, An Experimental study on Hypothyroid using Rotation Forest ,International Journal of Data Mining & Knowledge Management Process(IJDKP),vol.4(.6),pp.36-37, 2014.
2. A. Upadhyay,S. Shukla and S. kumar, Empirical Comparison by data mining classification algorithms(C4.5 & C5.0) for thyroid cancer data set, International Journal of Computer Science & Communication Networks, vol. 3(1,),pp.64-68.
3. N. Sigh and A. Jindal, A Segmentation Method and Comparison of Classification Methods for Thyroid Ultrasound Images, International Journal of Computer Applications, Vol. 50(11), 2012.
4. M. C. Frates, C. Benson, J. Charbonneau and S. Edmund ., "Management of Thyroid Nodules Detected at US: Society of Radiologists in US consensus", Conference statement management of thyroid nodules detected at US, Vol. 237(3), 2005.
5. Thyroid Data Set [online]. Available: https://archive.ics.uci.edu/ml/dataset s/Thyroid Disease, (Browsing Dec 2016).
6. H. Jiawei, K. Micheline, P. Jian, Data Mining Concepts and Techniques, Morgan Kaufmann, 2006.
7. D. Kerana Hanirex and K. P. Kaliyamurthie , Multi-Classification Approach For Detecting Thyroid Attacks",International journal of Pharma and Bio sciences, vol.4(3),pp.1246-1251,2013.
8. D. Lavanya & .K. Usha Rani, Performance Evalution of Decision Tree Classifiers on Medical Datasets, International Journal of Computer Application, vol.26(4),2011.
9. R. Parimala and R. Nallaswamy, A Study of Spam e-mail Classification using Feature Selection Package. Global Journal of Computer Science and Technology, Vol. 11, 2011.
10. K. J. Cios, W. W. Pedrycz , and R. W. Swiniarski , Data Mining Methods for Knowledge Discovery. Kluwer Academic Publishers, 3rd ed., 1998.
11. M. Pal, Ensemble Learning with Decision Tree for Remote Sensing Classification.World Academy of Science, Engineering and Technology. 36: 258-260, 2007.
12. S. Pandey,A. Tiwari , A. K. Shrivas and V. Sharma, Thyroid Classification using Ensemble Model with Feature Selection, International Journal of Computer Science and Information Technologies, Vol. 6 (3) , pp. 2395-2398,2015.
13. K. Geetha and Baboo C. S. Santosh, Efficient Thyroid Disease Classification Using Differential Evolution With SVM, Journal Of Theoretical And Applied Information Technology , Vol.88(3),410-420,2016.
14. K. Rajam, A Survey on Diagnosis of Thyroid Disease Using Data Mining Techniques, International Journal of Computer Science and Mobile Computing, Vol. 5(5), pp.354–358, 2016.
15. Source : Help File of Rapid Miner (Browsing date: Feb. 2016).